VISUALIZING DISTRIBUTIONS FROM MULTI-RETURN LIDAR DATA TO UNDERSTAND FOREST STRUCTURE

David Kao[1], Marc Kramer[1], Alison Love[2], Jennifer Dungan[1], and Alex Pang[2]
NASA Ames Research Center ({David.L.Kao,Jennifer.L.Dungan}@nasa.gov, mkramer@mail.arc.nasa.gov)
[2]Computer Science Department, University of California, Santa Cruz ({alison,pang}@soe.ucsc.edu)

Abstract

Spatially distributed probability density functions (pdfs) are becoming relevant to the Earth scientists and ecologists because of stochastic models and new sensors that provide numerous realizations or data points per unit area. One source of these data is from multi-return airborne lidar, a type of laser that records multiple returns for each pulse of light sent towards the ground. Data from multi-return lidar is a vital tool for understanding the structure of forest canopies over large extents. This paper suggests visualization tools to allow scientists to rapidly explore, interpret and discover characteristic distributions within the entire spatial field. The major contribution of this work is a paradigm shift that allows ecologists to think of and analyze their data in terms of full distributions, including their modality and shape, not just summary statistics. The tools allow the scientists to depart from traditional parametric statistical analyses and to associate multimodal distribution characteristics to forest structures. Examples are given using data from southeast Alaska.

I. INTRODUCTION

The development of remote sensing has created the routine need for analyzing large quantities of multidimensional and multivariate data. The dimensions are spatial, spectral and sometimes temporal. Spectral data has also been considered multivariate, where each waveband represents a variable and coincident data on variables measured by some other means may also need to be concomitantaly analyzed. In this paper we address a different aspect of some remote sensing data sets, their multi-valued nature. That is, we suggest that there are many situations where there exist multiple values of a single variable for each grid cell or spatial unit. Complete information on these multiple values is contained in the probability density function (pdf) of the variable at each location. These data sets are common products from remotely sensed images data and also from geophysical simulation. They are difficult to visualize because there are at least four axes of information.

Lidar (Light Detection And Ranging) is a remote sensing technology that yields multivalued datasets. When lidar sensors are used to measure vegetated surfaces such as forests, they can yield a detailed understanding of the canopy structure across an entire study area rather than at a few select plots. Distribution data from raw multi-return lidar data from forests provides information on forest structure, tree size and density [3]. Forest plots recovering from natural disturbance tend to have unimodal distributions of stem sizes and canopy heights with low standard deviations, whereas older, less disturbed forest plots tend to have multimodal distributions [6].

Previous work with lidar data has relied on statistical summaries that attempt to characterize each distribution with a small set of descriptors. The summaries reduce the dimensionality of the dataset and make visualization straightforward. This approach fails when the distributions are nonparametric or, especially, multimodal. We can expect many

distributions from lidar data to be multimodal when there exist distinct understorey and overstorey tiers or other recognizable vertical strata.

In this paper, we propose an approach to visualizing lidar data that allows exploration, query and comparison of distributions. The approach should increase opportunities to query multivalued data in new ways in order to better understand the distributions of geophysical and ecological phenomena, both at single locations and across the spatial domain. Interpretation can be gained from field reconnaissance, expert knowledge or ancillary information.

II. BACKGROUND

The challenge to visualizing spatially explicit, multimodal distributions is the four dimensional nature of the problem. To consider probability density functions (pdfs) over space, two dimensions are the orthogonal spatial dimensions, a third is the variable scale (in this case the height scale given by lidar) and the fourth is the frequency scale. Previously, we have reported on techniques for visualizing 4D spatial distribution data sets [4] using parametric statistics. That is, the pdf at every cell is characterized by a few statistical parameters such as mean, standard deviation, skewness, etc. and visualized. When some of the pdfs have multimodal distributions, statistical summaries are not sufficient. To address this, we have also used shape-based descriptors [5]. The basic idea here is to describe the shape of a distribution using the number of modes, the location of the modes, the width and height of each mode, etc. This descriptive information is then mapped to visual parameters. We demonstrate how that approach can be used with lidar data in Section V.

Previous efforts to visualize lidar data [2] presented ways in which a user can navigate through forest lidar data sets within a virtual environment. This is essentially the creation of a digital elevation model of the canopy top. Unlike this approach, our approach looks at aggregated multiple lidar returns. Therefore the data at each cell location is actually a collection of height values. In this study, we visualized distributions from 0.1 hectare cells, the size of field plots for which forest stand measures exist. The techniques developed in [4,5] are brought to bear upon this problem.

III. DATA

Forest canopy height distribution data were collected using a digital airborne topographic imaging system (DATIS-2; 3-Di Technology, MD, USA), a small-footprint lidar. The sensor is capable of retrieving multiple (up to 5) returns of elevation and intensity for every shot as it passes through a forested canopy. Over wooded terrain, the first return measures forest canopy height, while the last return measures ground elevation. DATIS-2 was flown in a Cessna 206 in May 2001 over High Island (approximately 500 hectares), located in the middle of the Alexander Archipelago. The data were initially collected at a density exceeding 2 shots per m^2. Raw data were processed into 81 measures of maximum forest canopy height for each 0.1 hectare cell across the island, resulting in 1800 0.1 grid cells with distribution data for each cell. The island is dominated by productive western hemlock (*Tsuga heterophylla* (Raf.) Sarg.) with scattered Sitka spruce (*Picea sitchensis* (Bong.) Carr.).

IV. ALGORITHMS

Prior to visualization, algorithms are applied to the raw data to estimate and characterize their distributions. In particular, density estimation is used to generate a probability density

function from the 81 heights at each grid cell, a peak hunting algorithm is used to find all the modes in the pdfs, and an operator is selected to allow distribution matching.

A. Density Estimation

For each grid cell in the field, there are multiple lidar returns, each with an associated height. These represent a sample of the full set of heights of all elements in the canopy. We use each sample to make an estimate of the "true" density, that is, the distribution of the full set of heights. One common density estimator, the histogram, does not produce a mathematically valid pdf and is very sensitive to the bin width used. There are many other estimators possible depending on the nature of the data [8]. In this application with lidar data, we selected a kernel estimator because it provides robust density estimation and is widely used.

B. Mode Finding

Given a distribution, there are a number of ways to characterize its modality, or how bumpy it is. For example, one may treat the distribution as a signal and apply Fourier analysis to extract the major frequency with the largest amplitude. However, one is still faced with the task of deciding what is a significant mode in frequency space. Alternatively, one may use Gaussian mixture models to fit the data as a weighted combination of Gaussian distributions. However, this approach requires significant a priori knowledge about the modality of the distribution in the first place. Instead, we use a descriptive approach that analyzes the shape of the distribution. We used the peak-hunting algorithm that we proposed in [5] to determine the number of peaks in each distribution and their respective positions.

C. Distribution Matching

There are a number of operators that can be applied as distance measures to compare distributions. The Kullback-Leibler (KL) distance, derived from information theory [1,7], can be used to compare distributions in order to find ones that have shapes that match a distribution of interest. The greater the KL distance, the less similar two distributions are. We set a threshold to control how similar we want the search results to be to a distribution of interest or target distribution. All the distributions with a distance less than the threshold to the target distribution will be accepted as similar distributions to the target distribution.

V. VISUALIZATION TOOLS

In the following subsections, we describe techniques designed to provide capabilities ranging from synoptic, general views of the full data set to more specific, localized and detailed query and display. The techniques allow extensions well beyond summary descriptors such as the quadratic mean, robust mean and chi-square or other “non-parametric" summaries or clustering algorithms. For each technique, we describe how it can assist the scientist in exploring distributions. We apply these tools to distributions of canopy height derived from the lidar data, focusing on several characteristic distributions that are of particular interest to the scientist. Collectively, these tools can be used effectively to analyze distribution data.

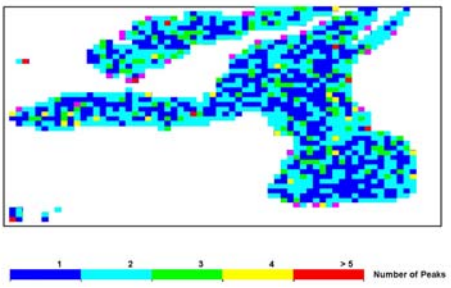

Figure 1: The spatial locations of unimodal and complex multimodal distributions in the High Island lidar data.

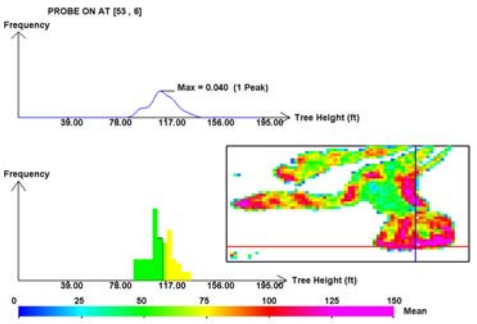

Figure 2: The kernel density estimate and the histogram at the probe position shown on the right image plane of the figure.

A. Displaying the modality of the distribution data

The first look at the spatially explicit distribution data shows the number of modes for each cell, calculated using the peak hunting algorithm described in [5] (Figure 1). This synoptic descriptor of all the data across the sample space provides the scientist with the first glimpse of new information related to the distribution. The modality of each cell gives some indication of patterns of multimodality. This is the first new information about the distribution that is not available through coarse statistical descriptors. This display helps answer questions such as what proportion of the data is unimodal or multimodal? Are the number of modes spatially clustered or concentrated in any one subregion of the field?

B. Interactive Data Probe

We have implemented an interactive data probe that allows the user to view the distribution of an individual cell at the current probe position set by the user. The interactive data probe is straightforward and useful for visualizing the pdf at any location in the field. It provides a per point basis query and shows the modality of the distribution. Only one density estimate is displayed at a given time (Figure 2). To begin gaining familiarity with the data the scientist can probe the forest data at different locations in order to have a good overall feel for the distributions in the study area. In addition, when viewing the distribution of a given cell, adjacent cells can be selected (through an up, down, left and right keyboard feature), thus allowing the scientist to visually traverse portions of the forest. This feature allows the scientist to view and relate distributions of particular forest regions of interest that s/he might already be familiar with through field reconnaissance or other ancillary data.

C. Mode Exploration

The modes of a distribution can be explored in a variety of ways using mode exploration tools. For all of these tools, the mode is computed using the peak hunting algorithm described in [5]. Meaningful mode exploration depends on using the proper density estimator, so that shapes in the distribution are real and not an artifact of low sampling density in the data. Conversely, it is important that all real information in the distribution be retained, so the smoothing function must minimize the loss of real information contained in the distribution. Our mode exploration tools comprised of: (1) mode query, (2) visualizing the distributions from the results of a mode query, and (3) visualizing the distributions and the spatial locations from the results of a mode query. Each of these processes is described in more detail in the following sections.

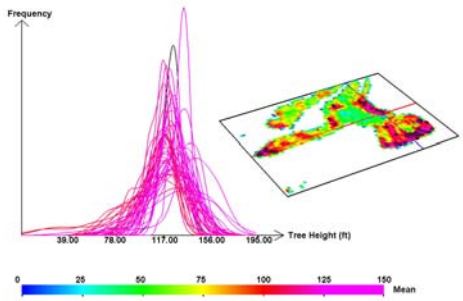

Figure 3: The image plane shows the mean canopy height for each cell in the field colored by six classes. Cells marked by black squares denote those locations found by the query: (1) mode between the heights of 117 and 194 feet, and (2) frequency above 0.05. The left graph shows the distributions of those pdfs matching the query.

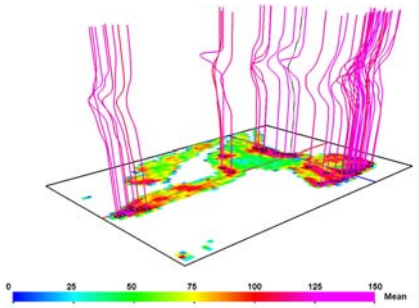

Figure 4: Same query results as shown in Figure 3 except that the pdfs are plotted directly above the grid cells. These pdf curves provide another visual cue of the distributions found by a mode query.

1) Mode query: These are queries that show the abundance, the spatial location, the spatial extent and the spatial pattern of distributions with a specific mode. For example, after using the interactive data probe and relating the distributions observed with field observations, the scientist became interested in finding all unimodal distributions with a mode between 117 and 194 feet. The mode query tool allows for visual identification of all those distributions that match this query (Figure 3). These locations are grid cells denoted with black squares. It is not surprising that the mean field of the distribution data at these grid cells are relatively high as indicated (in the red and magenta color range in Figure 3).

2) Visualizing the distributions from a mode query: From a mode query, there may be tens or hundreds of grid cells that match a specified mode criterion. The visualization challenge here is how to display all of these density estimates effectively, so that the scientists can begin to explore the shapes and diversity of the distributions identified through the query. Furthermore, these similarities and differences need to be highlighted for analysis e.g. for those pdfs that are very different, the information about which grid cells these pdfs represent should be shown/highlighted. Similarly, the grid cells of those conforming pdfs (pdfs that are similar) should be clustered or colored in the same group.

The most common approach to view several pdfs is to simply plot them side by side for visual comparison. This can be done by plotting a set of pdfs, or as many pdfs that can possibly fit on the screen in multiple windows. If the query only found a few pdfs, then this method is ideal and effective for comparing these pdfs. However, if there were tens or hundreds of pdfs found by the query, the user would need to view so many graphs as to make this method impractical.

Another simple approach is to plot all of the pdfs in one single graph, giving the scientist a visual comparison of these pdfs. This method is only useful, however, if the scientist is interested in determining whether there are any pdfs that differ significantly from others. The scientist would be able to see the overall shapes of these pdfs using this approach. However, for more detailed comparisons of pdfs, this method would not be suitable since it is most likely that many pdfs would overlap in the graph which makes it difficult to distinguish the details. In Figure 3, a graph of the distributions of those pdfs matched a

mode query is shown to the left. By displaying all of these pdfs in the same graph, the overall shape of these distributions can be seen to be very close.

Visualizing the distributions identified through a mode query not only allows the scientist to inspect them, it also allows the scientist to look for the following features: (1) outliers (how are outliers shaped, how many modes do they have, etc.), (2) trends (how are most of the distributions shaped, how many modes do they have), (3) diversity (how different are they from one another), (4) homogeneity (how similar), and (5) modality.

3) Visualizing the distributions and their spatial locations from a mode query: At this point, a scientist using the mode query tool has an idea of the locations and a graph of all the distributions that match the query. One source of information that is missing from the previous approach, shown in Figure 3, is that we do not know which grid cell the pdfs correspond to. In Figure 4, the same pdfs from Figure 3 are plotted right above their grid cells. We found this technique to be effective also for revealing the pdfs found by the query. By plotting the pdfs right above the corresponding grid cells, we can easily see the spatial locations of the matching pdfs. Note that the pdfs are drawn such that the density estimates are plotted along the axis perpendicular to the image plane. We construct a pdf curve for each grid cell found by the mode query. A pdf curve is created by horizontally displacing points along a vertical line by the magnitude of the density estimate. The height of the pdf curve is determined by the number of evaluation points of the density estimate. In our example, 150 evaluation points are used. The color of each pdf curve presents the mean tree height of the distribution data at the corresponding grid cell.

D. Distribution Exploration

Distribution exploration is performed by distribution matching and visualizing similar distribution shapes. This allows scientist to identify all distributions that are similar in their entirety rather than in just a mode. Our tool allows the user to be more restrictive or more relaxed in the specificity of finding “like” distributions and allows all distributions to be ranked in terms of their similarity to the specified pattern. For example, matching could be restricted to certain data range, or only when the frequency is above a certain threshold. Likewise, matching could be relaxed by lowering acceptance threshold or using more liberal similarity metrics. Since density estimates vary in their quality, the ability to relax or restrict the definitions of similarity with the query tool allows user flexibility in identifying a range of like distributions and their spatial locations. Through using the interactive data probe, the scientist is able to “visit” portions of the forest he was already familiar with (through field reconnaissance). Using the distribution matching tool, all distributions that are similar in their entirety to a specific distribution of interest were identified using contour lines as illustrated in Figure 5. Identifying similar or matching distributions can be a powerful way to perform hypothesis testing, guide additional field work, and generate new data products of interest.

Once the contours lines are generated from the results of the distribution matching tool, an additional visualization tool provides yet another way of studying subtle differences in the matching distributions. This tool constructs color-mapped characteristic distribution surfaces to depict the variations of the pdfs along the contour lines. For each grid cell along a contour line, a vertical line is plotted right above the corresponding grid cell. Then, a surface mesh is formed by connecting vertical lines from the adjacent points along the contour line. The surface mesh is colored by the density estimates. Since there are usually several contour lines, our tool would generate several disjoint characteristic distribution surfaces. As with the pdf curves shown in Figure 4, the height of the surfaces is determined

by the number of evaluation points of the density estimate. Figure 6 shows the characteristic distribution surfaces of the matching distributions. As with the visualization tools provided in the modality exploration, this visualization tool provides even further refinement of relative homogeneity, heterogeneity and associated possible spatial patterning of the characteristic distributions.

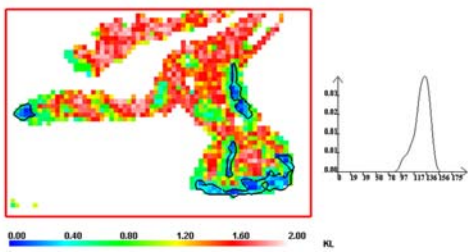

Figure 5: Using the distribution matching tool, the scientist found all distributions that are similar to one (graph shown on the right) found to be recovering from a recent disturbance event.

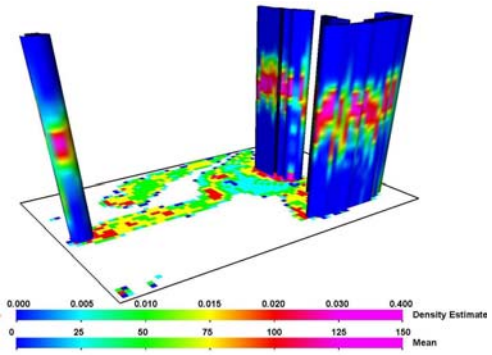

Figure 6: Characteristic distribution surfaces for the contours lines shown in Figure 5. The surfaces are colored by the density estimates. The pdfs along the contours are mostly unimodal as indicated by the central magenta color band that runs across the middle.

VI. DISCUSSION

The utility of each of the technique discussed in this paper fundamentally depends on the selection of an appropriate density estimator. The estimator determines how the data are smoothed and how modes are defined. Each estimator is different, and may be well-suited for one type of data but not another. The kernel estimator used in this application with lidar data, for example, may have smoothed possibly interesting features in the data. The appropriateness of a given estimator depends partly on the number of raw data values per grid cell. In general, the larger the number of raw values, the more robust a given esimator will be. The precision of the data can also affect the size of the kernel used and the consequent smoothing of the distribution.

A key feature of these tools is their flexibility. Software that gives the scientist a choice of estimator and the ability to specify the parameters used in estimation will allow the accomodation of diverse data sets and exploratory data analysis. The kernel estimator we used in this study was selected because it is a robust, widely-used estimation technique but many other choices are possible.

Once an estimator is selected, the identification of modes is also not completely deterministic. Small bumps may be of little or no interest to the scientist, so what constitutes a mode in the display and query of modality can be user defined. Matching entire distributions is also a user-defined process. Success depends on increasing or decreasing the specificity of the distribution matching algorithm and having some meaningful criteria for doing so. Also, the distribution matching algorithm used is important. In this paper we used the KL distance, but others are possible. Ultimately we envision a user-selection capability, so that various algorithms can be employed and their output assessed.

VII. CONCLUDING REMARKS

Overall, our visualization tools provide new ways to query, visualize and compare distributions. The key contributions of this work are (1) automated ways to process forest canopy distributions derived from lidar data and (2) improved interactive access to lidar distributions, allowing the scientist to form and test hypotheses about horizontal and vertical structure in forests.

Though the application described in this paper deals specifically with multi-return lidar data, our tools can be easily be used with distribution data sets from other applications. There are several open research problems in visualizing spatially varying distribution data sets, including the extension to distribution data that are sampled in a 3D domain and the extension to distribution data on more than one variable at a time.

ACKNOWLEDGEMENTS

This work is supported in part by the NASA Intelligent Systems Program Cooperative Agreement NCC2-1260 and NSF ACI-9908881. Additional support was provided to the second author through a National Research Council (NRC) postdoctoral research fellowship. We would like to thank Chris Hlavka and Michael Gerald-Yamasaki for their helpful comments and Anna Chen, Newton Der, Jose Renteria, Wei Shen, and Bing Zhang for help with programming and data preparation.

REFERENCES

[1] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
[2] N. T. Eggleston, M. Watson, D. L. Evans, R. J. Moorhead, and J. W. McCombs, II. Visualization of airborne multiple-return LIDAR imagery from a forested landscape. In *Second International Conference on Geospatial Information in Agriculture and Forestry, 10-12 January 2000*, Lake Buena Vista, Florida, 2000.
[3] J. Hyyppa, O. Kelle, M. Lehikoinen, and M. Inkinen. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. *IEEE Transactions on Geoscience and Remote Sensing*, 39:969 –975, 2001.
[4] D. Kao, J. Dungan, and A. Pang. Visualizing 2D probability distributions from EOS satellite image-derived data sets: A case study. In *Proceedings of Visualization '01*, 457–460, 2001.
[5] D. Kao, A. Luo, J. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *Proceedings of the 6th International Conference on Information Visualization '02*, pages 219–225. IEEE Computer Society, 2002.
[6] M. G. Kramer, A. J. Hansen, M. Taper, and E. Kissinger. Abiotic controls on windthrow and forest dynamics in a coastal temperate rainforest, Kuiu Island, southeast Alaska. *Ecology*, 82:2749–2768, 2001.
[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 1951.
[8] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.